CP^2: Copy-Paste Contrastive Pretraining for Semantic Segmentation

Feng Wang[1], Huiyu Wang[2], Chen Wei[2], Alan Yuille[2], and Wei Shen[3⋆]

[1] Department of Automation, Tsinghua University
[2] Department of Computer Science, Johns Hopkins University
[3] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

Motivation

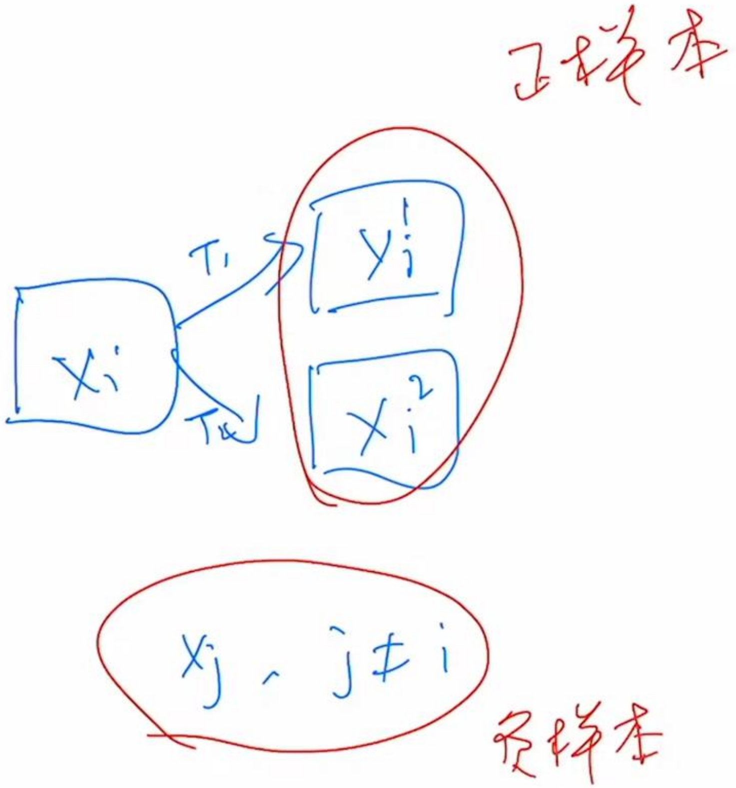
- Current image-level pretraining to downstream dense prediction paradigm is not ideal
- Existing constrastive learning model may over-fit to learning image-level representation and neglect pixel-level variances
- Arch. misalignment:
    - Sem. seg. requires small out stride and a large atrous rate (not in backbone)
    - Randomly initializaiton for seg. head could negatively affect trained bone

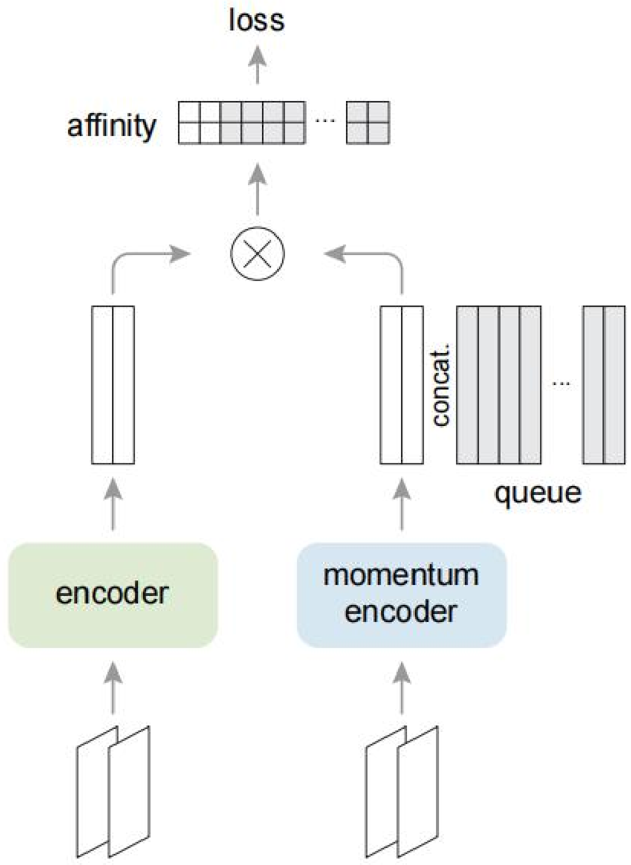
Contribution

- Novel Self-pretraining Method (**C**opy-**P**aste **C**onstastive **P**retraining)
  - Address arch. misalignment
  - Endow net with perception of spatially varying information
- Quick Tuning protocol
- +2.7% mIoU on PASCAL VOC 2012 than baseline (MoCo v2)

Review -- MoCo

- Instance Discrimination (个体判别)

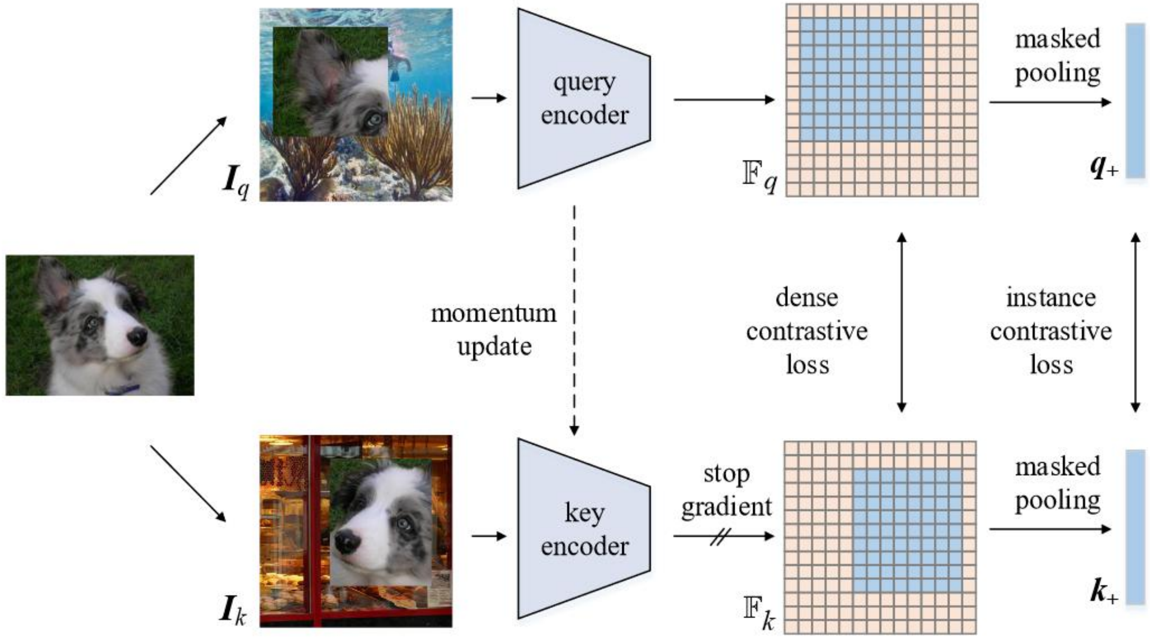




(b) Momentum Contrast

Method

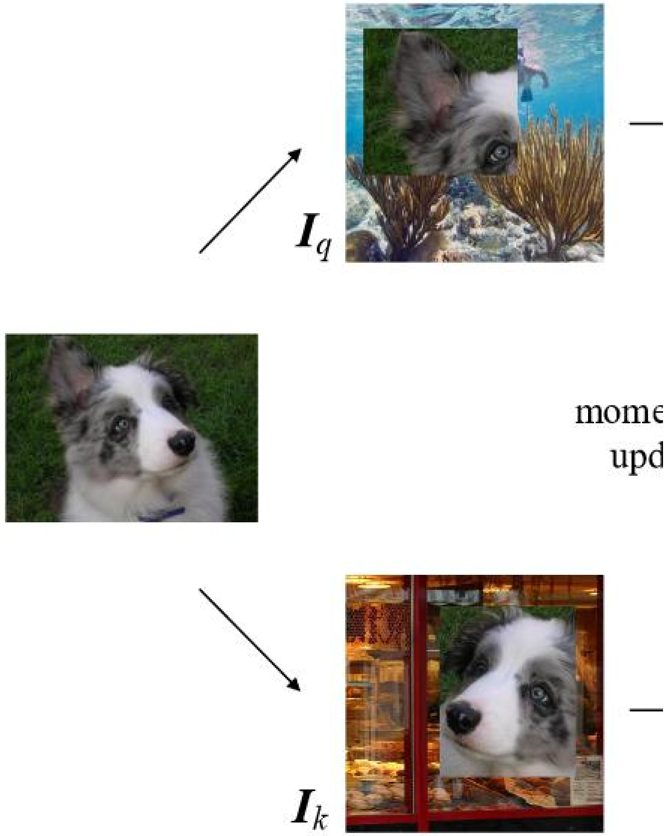
- Rand crop and paste to diff. backgrounds
- Two targets:
  - Get FG from BG
  - Recognize Ims with same FG



Method

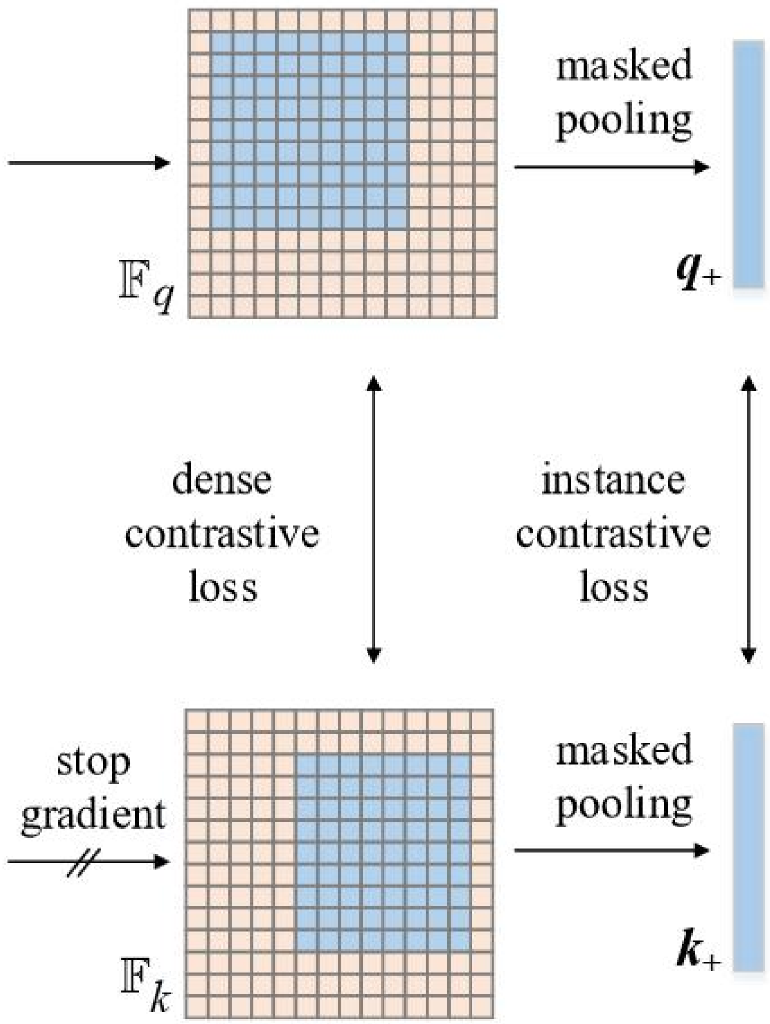
- Compose Images

$$\boldsymbol{I}_q = \boldsymbol{I}_q^{fore} \odot \boldsymbol{M}_q + \boldsymbol{I}_q^{back} \odot (1 - \boldsymbol{M}_q)\,,$$
$$\boldsymbol{I}_k = \boldsymbol{I}_k^{fore} \odot \boldsymbol{M}_k + \boldsymbol{I}_k^{back} \odot (1 - \boldsymbol{M}_k)\,,$$



Method

- Constrastive objectives
- -- Dense Constrastive
  - Diff. FG & BG
- -- Instance Constrastive
  - Keep global, instance-level representations

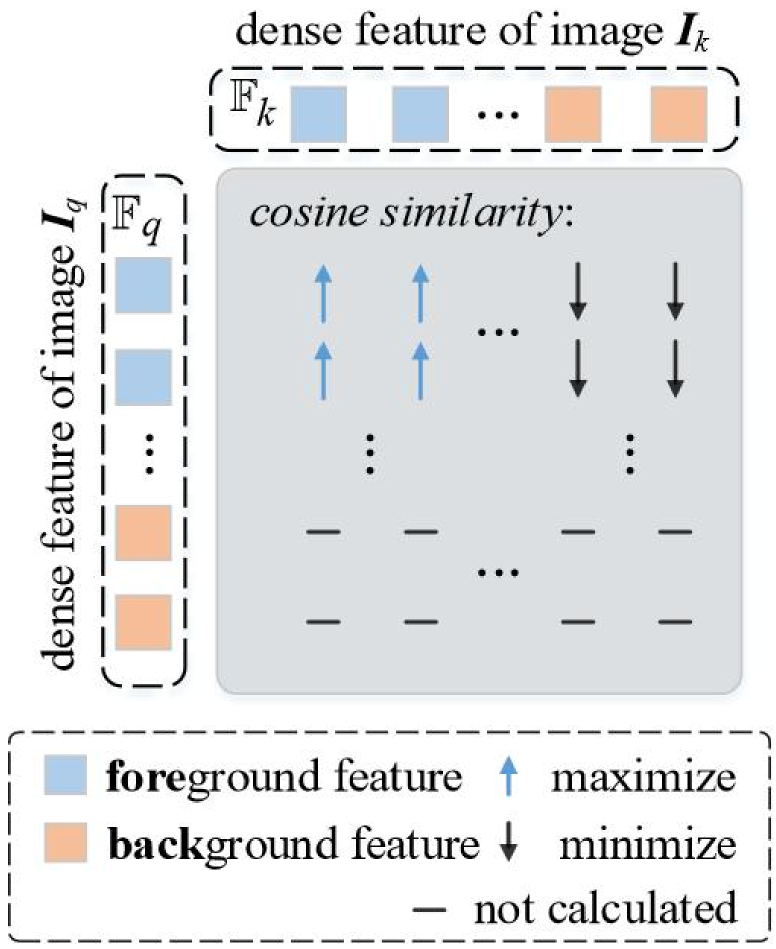


Method

- Dense Loss

$$\mathcal{L}_{dense} = -\frac{1}{|\mathbb{F}_q^+||\mathbb{F}_k^+|} \sum_{\forall \boldsymbol{f}_q^+ \in \mathbb{F}_q^+, \forall \boldsymbol{f}_k^+ \in \mathbb{F}_k^+} \log \frac{\exp\left(\boldsymbol{f}_q^+ \cdot \boldsymbol{f}_k^+ / \tau_{dense}\right)}{\sum_{\forall \boldsymbol{f}_k \in \mathbb{F}_k} \exp\left(\boldsymbol{f}_q^+ \cdot \boldsymbol{f}_k / \tau_{dense}\right)},$$

$$\boxed{\boldsymbol{f}_k^+ \in \mathbb{F}_k^+ \subset \mathbb{F}_k.}$$



Method

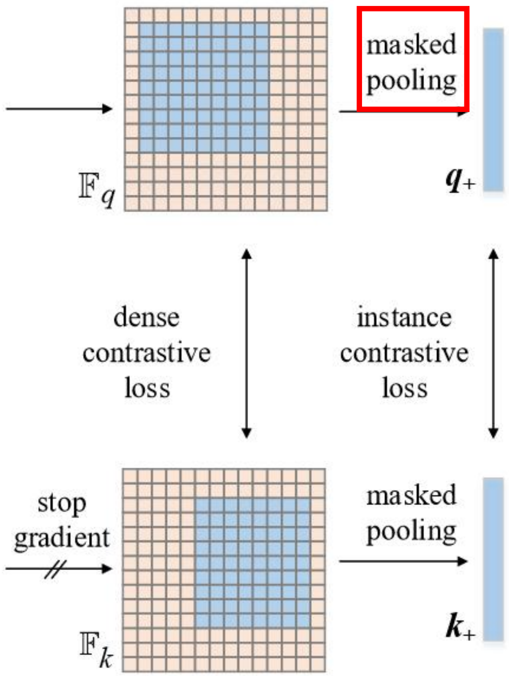
- Instance Loss
  - Distinguish positive key from **a memory bank** of negative keys
  - Use normalized masked averaging of only FG features

  instead of GAP

$$\mathcal{L}_{ins} = -\log \frac{\exp(\boldsymbol{q}_+ \cdot \boldsymbol{k}_+/\tau_{ins})}{\exp(\boldsymbol{q}_+ \cdot \boldsymbol{k}_+/\tau_{ins}) + \sum_{n=1}^{N} \exp(\boldsymbol{q}_+ \cdot \boldsymbol{k}_n/\tau_{ins})},$$

where $\boldsymbol{q}_+$, $\boldsymbol{k}_+$ are normalized masked averaging of $\mathbb{F}_q^+$ and $\mathbb{F}_k^+$:

$$\boldsymbol{q}_+ = \frac{\sum_{\forall \boldsymbol{f}_q^+ \in \mathbb{F}_q^+} \boldsymbol{f}_q^+}{|| \sum_{\forall \boldsymbol{f}_q^+ \in \mathbb{F}_q^+} \boldsymbol{f}_q^+ ||_2}, \quad \boldsymbol{k}_+ = \frac{\sum_{\forall \boldsymbol{f}_k^+ \in \mathbb{F}_k^+} \boldsymbol{f}_k^+}{|| \sum_{\forall \boldsymbol{f}_k^+ \in \mathbb{F}_k^+} \boldsymbol{f}_k^+ ||_2}.$$



$$\boxed{\mathcal{L} = \mathcal{L}_{ins} + \alpha \mathcal{L}_{dense},}$$

Method

- Arch.
  - Make some adaption on ResNet50
  - Compatible with most seg. heads (DeepLab v3 by default)
- Quick Tuning
  - Initialize backbone with available backbones
  - Initialize seg. head with random parameters
  - Use CP^2 training method for **just a few epochs**

Experiments

- Datasets
  - Pretrain: ImageNet
  - Segment: VOC, ADE20K, Cityscapes
- Seg. Head
  - DeepLab v3 ASPP (default setting)
  - CP^2 head: 2Layer 512Channel 1*1 + ReLu + 128C 1*1
  - FCN Head

Experiments

Table 1: **Evaluation results (mIoU) with DeepLab v3 segmentation head**. QT denotes Quick Tuning with CP^2 initialized by a MoCo v2 pre-trained backbone. Our results are marked in gray. The best results are **bolded**. Epochs that are consumed by the initialization model are de-emphasized.

method	backbone	epoch	PASCAL	Cityscapes	ADE20k
supervised	ResNet-50	-	76.0	76.3	39.5
MoCo [28]	ResNet-50	200	73.2	75.8	38.6
SimCLR [10]	ResNet-50	1000	77.3	76.5	40.1
BYOL [26]	ResNet-50	300	77.4	76.5	40.2
InfoMin [41]	ResNet-50	800	77.2	76.5	39.6
InsLoc [46]	ResNet-50	400	75.6	76.3	40.3
DetCon [30]	ResNet-50	1000	78.1	77.1	40.6
PixPro [45]	ResNet-50	400	77.5	76.6	40.3
MoCo v2 [12]	ResNet-50	200	74.9	76.2	39.2
CP^2	ResNet-50	200	77.6	77.3	40.5
CP^2 QT r.200	ResNet-50	200+20	76.5	77.2	40.7
MoCo v2 [12]	ResNet-50	800	77.2	76.4	39.7
CP^2 QT r.800	ResNet-50	800+20	**78.6**	**77.4**	**41.3**
MoCo v2 [12]	ViT-S/16	300	78.8	77.2	41.3
CP^2 QT v.300	ViT-S/16	300+20	**79.5**	**77.6**	**42.2**

Table 2: **Evaluation results (mIoU) with FCN head**. QT denotes Quick Tuning with CP^2 initialized by a MoCo v2 pre-trained backbone. Our results are marked in gray. The best results are **bolded**. Epochs that are consumed by the initialization model are de-emphasized.

method	backbone	epoch	PASCAL	Cityscapes	ADE20k
supervised	ResNet-50	-	73.7	75.8	37.4
MoCo v2 [12]	ResNet-50	200	74.4	75.8	37.4
CP^2	ResNet-50	200	75.4	76.4	38.4
CP^2 QT r.200	ResNet-50	200+20	75.2	76.4	38.0
MoCo v2 [12]	ResNet-50	800	74.8	75.9	37.9
CP^2 QT r.800	ResNet-50	800+20	**75.7**	**76.5**	**39.2**
MoCo v2 [12]	ViT-S/16	300	77.7	76.6	40.4
CP^2 QT v.300	ViT-S/16	300+20	**78.6**	**77.0**	**41.2**

Experiments

- Ablation
  - Seg head during pretraining & Dense Loss
  - Copy-paste style (pixel-wise, patch-wise and rec.-wise)
  - Training schedule
  - Hyper-parameters

Experiments

Table 3: **Ablation study of segmentation head pretraining** on PASCAL VOC. The results are based on ASPP segmentation head. We use Quick Tuning for CP^2 in the settings of (ResNet-50, 800 epochs) and (ViT-S/16, 300 epochs).

mode	backbone	head	mIoU
ResNet-50, 200 epochs	MoCo v2	random	74.9
	CP^2	random	76.3 (+1.4)
	CP^2	**CP**^{2}	**77.6** (+**2.7**)
ResNet-50, 800 epochs	MoCo v2	random	77.2
	CP^2 QT	random	78.2 (+1.0)
	CP^2 QT	**CP**^{2} QT	**78.6** (+**1.4**)
ViT-S/16, 300 epochs	MoCo v2	random	78.8
	CP^2 QT	random	79.3 (+0.5)
	CP^2 QT	**CP**^{2} QT	**79.5** (+**0.7**)

Experiments

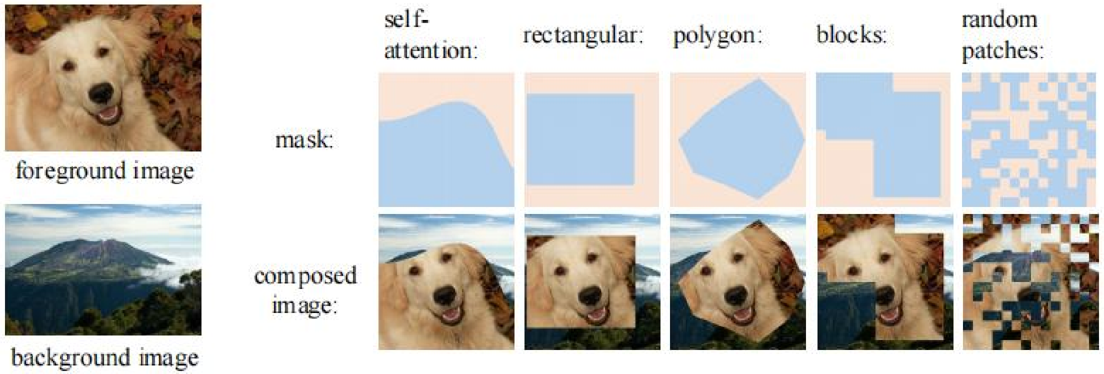


Fig. 4: **Examples of masking strategies and composed images**. The self-attention mask (DINO mask) is smoothed by Gaussian blur.

Table 4: **Evaluation results of foreground-background masks** on PASCAL VOC. Note that for the full mask, the models are trained without dense contrastive loss. Our default setting is marked in gray.

mode	random	mIoU	
		ResNet-50	ViT-S/16
baseline MoCo v2	-	77.2	78.8
no copy-paste	-	77.6	78.9
DINO self-attention mask [6]	✗	77.9	79.3
rectangular mask	✓	**78.6**	**79.5**
polygon mask	✓	78.1	79.0
random blocks	✓	77.3	78.7
random patches	✓	75.3	78.9

Experiments

Table 5: **Evaluation results of hyper-parameter search** on PASCAL VOC. The results are based on ResNet50-ASPP models, where the base backbone is loaded from the MoCo v2 pretrained ResNet50 for 800 epochs. Our default setting is marked in gray. The best results are **bolded**.

(a) **loss weight and temperature**

weight	temperature(τ_{dense}) 2	1	0.5	0.2
10	77.4	77.0	76.9	77.2
1	77.3	77.9	77.3	77.4
0.5	77.2	78.0	77.3	77.1
0.2	76.9	**78.6**	77.3	76.7
0.1	76.0	77.7	77.5	75.8

(b) **Quick Tuning epochs**

epoch	mIoU
0	77.2
10	77.7 (+0.5)
20	**78.6 (+1.4)**
40	78.7 (+1.5)

$$\mathcal{L}_{dense} = -\frac{1}{|\mathbb{F}_q^+||\mathbb{F}_k^+|} \sum_{\forall \boldsymbol{f}_q^+ \in \mathbb{F}_q^+, \forall \boldsymbol{f}_k^+ \in \mathbb{F}_k^+} \log \frac{\exp\left(\boldsymbol{f}_q^+ \cdot \boldsymbol{f}_k^+ / \tau_{dense}\right)}{\sum_{\forall \boldsymbol{f}_k \in \mathbb{F}_k} \exp\left(\boldsymbol{f}_q^+ \cdot \boldsymbol{f}_k / \tau_{dense}\right)},$$

Summary

- Constrastive loss for segmentation ?
- How about using this as backbone ?